Automatic morphological analysis of Basque

I. Alegria, X. Artola, K. Sarasola, M. Urkia

Automatic morphological analysis of Basque

Iñaki Alegria, Xabier Artola, Kepa Sarasola and Miriam Urkia
Basque Country University and UZEI

1 Introduction

The two-level model of computational morphology was proposed by Koskenniemi (1983) and has found widespread acceptance due mostly to its general applicability, declarativeness of rules and clear separation of linguistic knowledge and program. The essential difference from generative phonology is that there are no intermediate states between lexical and surface representations. Word recognition is reduced to finding valid lexical representations which correspond to a given surface form. Inversely, generation proceeds from a known lexical representation and searches for surface representations corresponding to it. The complexity of the model is studied in depth in (Barton, 85) who concludes that the complexity of a language has no significant effects on the speed of analysis or synthesis.

The two-level model of morphology has become the most popular formalism for highly inflected and agglutinative languages (Antworth, 90) (Sproat, 92) (Oflazer, 94). The two-level system is based on two main components —see Sproat (1992):

- A lexicon where the morphemes (lemmas and affixes) and the possible links among them (morphotactics) are defined. The lexicon is divided into different sublexicons and each lexicon entry specifies its morphotactical information by means of a continuation class which is a set of sublexicons. Combining sublexicons (nodes) and continuation classes (arcs) the graph of morphotactics is defined.
- A set of rules which controls the mapping between the lexical level and the surface level due to the morphonological transformations (morphophonemics). There are four kind of rules: context restriction rules "=>" (lexical character may be realized as the lexical one in the given context), surface coercion rules "<=" (lexical character must be realized as the lexical one in the given context), composite rules "<=>" (lexical character must be realized as the lexical one in the given context and this change is licit only in this context) and exclusion rules (lexical character may not be realized as the lexical one in the given context). The rules are independent from the morphotactics.

The rules are compiled into transducers, so it is possible to apply the system for both analysis and generation. PC-Kimmo (Antworth, 90) is a freely available software tool which is useful to experiment with this formalism. Different flavours of two-level morphology have been developed, most of them changing the continuation class based morphotactics by unification based mechanisms (for instance Ritchie *et al.*, 92). At Xerox

have been developed the lexical transducers (Karttunen, 94) (Alegria *et al.*, 95) which improve the speed and expressivity of the two-level formalism.

We have developed our own implementation of the two-level model with slight variations —an extension for continuation class specifications in order to deal with long-distance dependencies, for instance—, and applied it to Basque (Agirre et al., 92). In order to deal with a wide variety of linguistic data and to be a support for other NLP applications, we have built a Lexical Database (LDBB). At present it contains 60,000 entries, each with its associated linguistic features (category, subcategory, case, number, etc.).

In order to increase the coverage and the robustness, the analyser has been designed in an incremental way and it consists of three main modules (see Fig. 1): the standard analyser, the analyser of linguistic variants —due to dialectal uses and competence errors—, and the analyser without lexicon which can recognize word-forms without having their lemmas in the lexicon. An important feature of the analyser is its homogeneity as the three different steps are based on two-level morphology, very different from ad-hoc solutions.

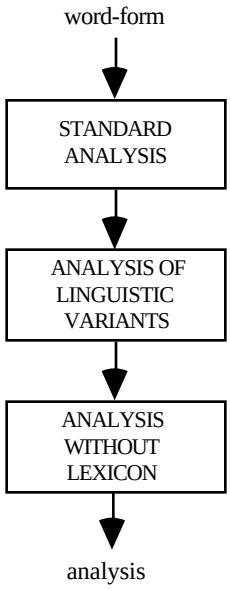


Fig. 1 Modules of the analyzer

This analyser is a basic tool for current and future work on automatic processing of Basque and its first two applications are a commercial spelling corrector named *Xuxen* and a general purpose lemmatizer/tagger (Aduriz et al., 95) named *EUSLEM*.

2. Brief Description of Basque Morphology

Basque is a preindoeuropean language with an unknown origin and quite different from the surrounding European languages

There are approximately 700.000 speakers and six dialects. The dialects are very distinct from other. In 1968 the Basque Academy of the Language decided to create the Standard Basque and it has been very well accepted, so the unified language, used today in TV, radio, school, university and so on, is only 27 years old. It means that there are lot of problems in the way of development. Quite descriptive works have been made; morphology, for instance, is quite well described and standardized, but there is still hard work to be done. The problem is the prescription, because a language in ways of standardization needs rules and decisions and, in fact, the Academy has been making some important decisions in the two last years.

These are some of the most important features of Basque:

- it is an agglutinative language; the determiner, the number and the declension case are appended to the last element of the phrase and always in this order. These information is valid for all the elements of the phrase. For instance, *semeArEN etxeAN* (in the house of the son):

seme	***A***	***r***	***EN***	***etxe***	***A***	***N***
noun	determiner	epenthetical	genitive	noun	determiner	inessive
(son)		element	case	(house)		case

- Basque has an only declension table, i.e., the 15 cases do not change, their morphemes are always added to the other elements, but is not like Latin, for instance, where there are five declension tables.
- As prepositional functions are realized by case suffixes inside word-forms, Basque presents a relatively high power to generate inflected word-forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. Moreover, while 77 of them are simple combinations of number, determination, and case marks, not capable of further inflection, the other 58 are word-forms ended with one of the two possible genitives (possessive and locative) or with a sequence composed of a case mark and a genitive mark. If the latter is the case, then by adding again the same set of morpheme combinations (135) to each one of those 58 forms a new, complete set of forms could be recursively generated. This kind of construction reveals a noun ellipsis inside a complex noun phrase and could be theoretically extended *ad infinitum*; however, in practice it is not usual to find more than two levels of this kind of recursion in a word-form. Related high power of generation is similar in most of the roots with declension but is higher with the adjectives where, due to the three cases of degree, the possible combinations are multiplied by 4.
- In Basque more than about morphology we can speak about morphosyntax. For instance, the case morpheme adds syntactic information inside the word-form.
- Gender does not exist in Basque; the only gender difference is in the allocutive verbs, where in familiar treatment it distinguishes male and female in the second singular person.

- The verb offers all the grammatical information. A verb form tells us who the subject is, the two objects, as well as the tense, aspect, etc.

  ex.: *daramazkiot* 'I take something (plural) to him/her', where
  d- (direct complement)
  -a- (present tense)
  -rama- (root 'take to')
  -zki- (plural)
  -o- (second complement, dative)
  -t (ergative mark, subject)
- The verb can be periphrastic or synthetic. The synthetic forms are used in the old verbs but it is not productive nowadays.
- We do not need always to explicit the grammatical person in a sentence, because it can be understood with the verb. The subject, for instance, always can be implicit.
- Ergative case exists in Basque. There are some theories about that. For some linguists it is an ergative language, for others it is non-accusative, but, in fact, ergativity exists.

  Depending on transitive/intransitive sentences, the case changes.

Ni etorri naiz		I came (intransitive)
Subject Absolutive		
Nik erosi dut liburua		I bought the book (transitive)
Subject Ergative	Direct compl. Absolutive	

  So, intransitive sentences have the subject in absolutive case, but in transitive sentences the absolutive goes to the direct complement and the subject takes the ergative.
- The order of sentence elements is free. Often the order change is related to the topic/focus.
- Word-formation is very productive in Basque. It is very usual to create new compounds as well as derivatives (prefixes and affixes are very normal, and infixes are almost in old forms, not used nowadays).

All these features have made the automatic treatment of our language difficult, especially because of the lack in theoretical studies. In our system, inflectional morphology of Basque has been completely described as we show later on, but the treatment of derivation and composition has not been exhaustive. Derivational morphology has been treated by lexicalized terms with the exception of the few cases where the generalisation was possible, always with the goal of avoiding overgeneration. Composition has been *discarded*[1] when the unit of treatment is longer than one word and, in the other cases, we have worked as in the derivation: when generalisation was possible (i.e. noun-noun case) we described it but otherwise only lexicalized terms are accepted.

---

[1] This treatment will take place in the process of lemmatization/tagging.

3 The Standard Morphological Processor

We have applied the two-level model defining the following elements (Agirre et al., 92) (Alegria, 95): lexicon, continuation classes and morphonological rules. Among the morphological phenomena handled by our system so far, we would like to emphasize the following: whole declension system —including place and person names, special declension of pronouns, adverbs, etc.—, graduation of adjectives, relational endings and prefixes for verb forms —finite and non-finite— and some frequent and productive cases of derivation and compounding.

In order to deal with a wide variety of linguistic data we have built a Lexical Database (LDBB). This database is both source and support for the lexicons needed in several applications, and has been designed with the objectives of being neutral in relation to linguistic formalisms, flexible, open and easy to use (Agirre et al., 95). The data base is permanently updated by linguists and exported as it is needed (see Fig. 2).

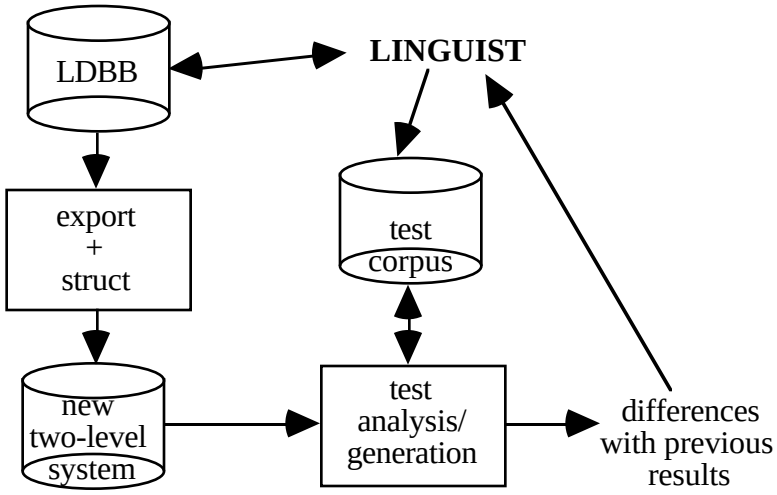


Fig. 2 Updating and using the lexical daba base

3.1 The Lexicon

Near to 60,000 entries have been defined corresponding to lemmas and affixes, grouped in 154 sublexicons. The most important information that can be associated to each entry in the lexicon are as follows:

- sublexicon where the entry is included
- canonical form and two-level form
- continuation class
- category and subcategory
- morphological information: number, determination, person, tense, ...
- morphosyntactic information: case, function, relational morphemes, ...
- additional information: source reference, example, frequency

Separated representation for homographs —in the main sublexicon, with the same or different continuation classes— has been made possible. Although this distinction is not

necessarily relevant to morphological analysis, future work on syntax and semantics has been taken into consideration. Table 1 shows the number of entries belonging to the most important sublexicons.

SUBLEXICON	ENTRIES
nouns	23.078
inflected verbs	7.387
adjectives	6.250
verbs	4.324
adverbs	1.714
initials	314
pronouns	308
other lemmas	1.957

Table 1 Number of entries in the lexicon

The two-level representation of the entries is not canonical because 18 *diacritics* are used to control the application of morphonological rules. The diacritics that we use are the following:

- R special r with a double sense: hard r at the end of the lemma and epenthetical r in the beginning of some suffixes.
- Q special r at the end of lemma which plays like a vowel with some suffixes.
- E epenthetical e
- N final n of old verbs which is lost with some suffixes
- M final n of suffixes which is lost sometimes
- \ special final n (with optional loss)
- A organic a of common lemmas
- # exceptional organic a (6 cases)
- & special final a of place names
- @ special final a of verbs that can be replaced by e
- ^ final character of verbal flexion meaning possible epenthetical a
- % final character of some place names meaning special declension
- : final character after consonant of some initials meaning vowel-like declension
- / final character after consonant of some initials meaning optional declension like consonant or vowel
- $ final character after vowel in inflected verbs to simulate final consonant
- + morpheme boundary

The surface characters are 30 —the 26 standard of the Latin alphabet, the ñ character, the hyphen, the point and the * symbol to mark capital letters— which added to the mentioned marks complete the lexical alphabet.

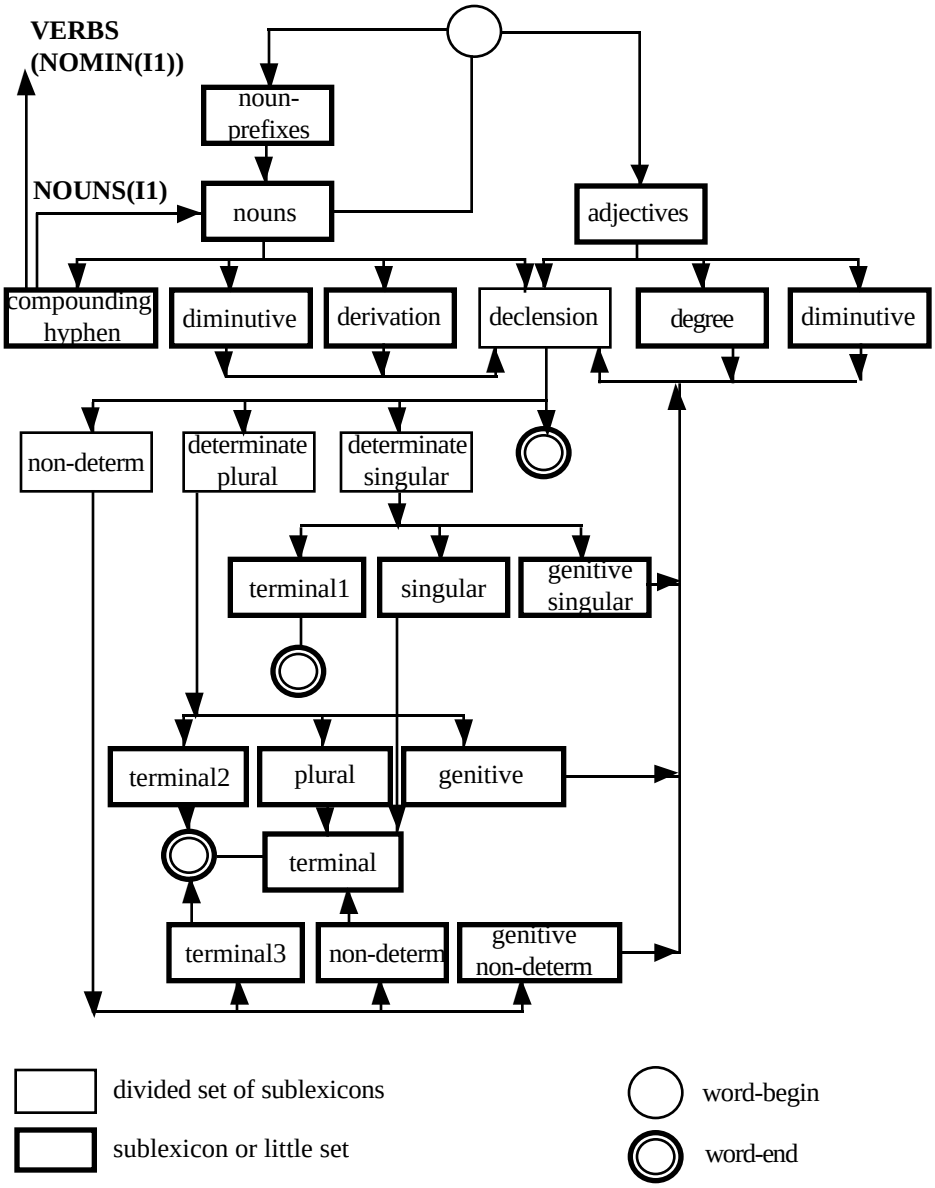


Fig. 3 Morphotactics for nouns and adjectives

3.2 The Morphotactics

Continuation classes are the basic elements of morphotactics if the original proposal of Koskenniemi is taken. They are groups of sublexicons that describe the morphotactics. Each entry of the lexicon has a continuation class assigned to it. All the continuation classes together define the morphotactics graph. Using this mechanism only linear links can be expressed, but long distance dependencies among morphemes can not be expressed. In order to deal with this problem, in our implementation we extended the semantics of the formalism, defining extended continuation classes as we show in the next section. Others

changes to the original morphotactical mechanism have been proposed by different authors (Bear, 86), (Ritchie et al., 87) (Trost 90).

Generalizations have not been always possible because we wanted to do an extensive definition avoiding overgeneration. For example, while with nouns and adjectives the assignment of a single continuation class to all of the elements of each category has been possible, adverbs, pronouns and verbs have required more particularized solutions. More than one hundred (130) different continuation classes have been defined. In Fig. 3 and Fig. 4 we show the main schemes of morphotactics for nouns, adjectives and verb infinitives.

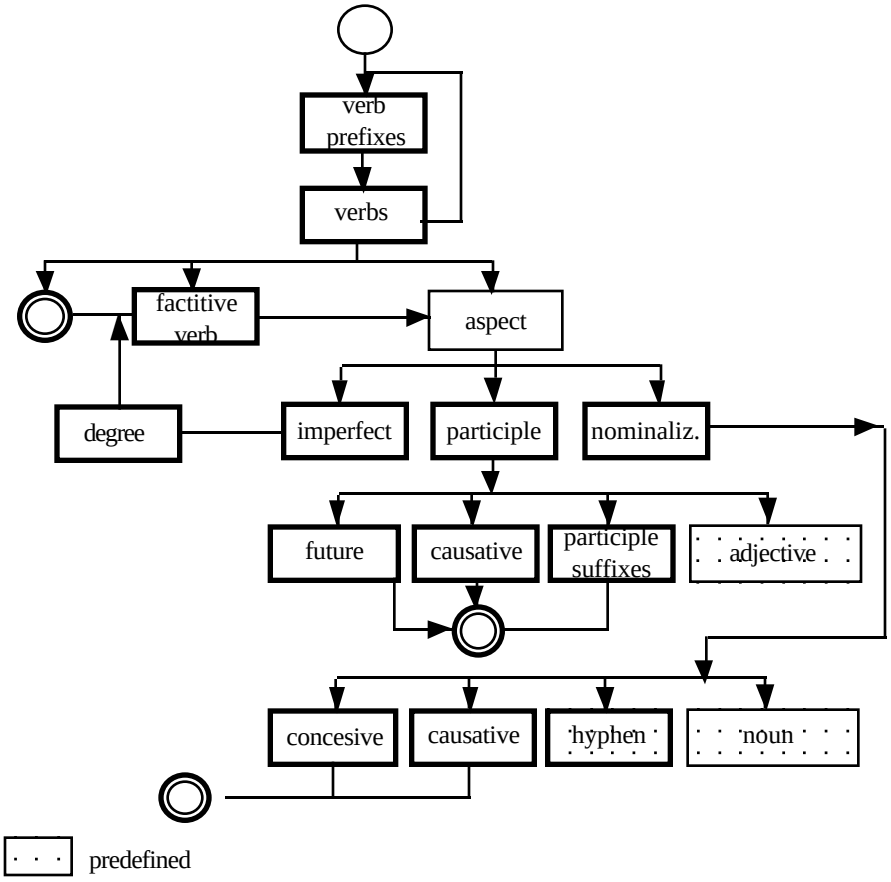


Fig. 4 Morphotactics for verbs

3.3 Solving Long-distance Dependencies

Up until now, the notation and concept of continuation classes have been used, in the authors' opinion this is the weakest point of the formalism. Specifically in dealing with the Basque auxiliary verb, many cases of long-distance dependencies that are not possible to express adequately in this way have been found. For instance in English, *en-*, *joy* and *-able* can be linked together, but it is not possible to link only *joy* and *-able*. So we say that the possibility of the suffix *-able* depends on the prefix *en-* that is not next to it.

Different solutions have been proposed to solve similar problems for several languages (Trost, 90). The solution that we have designed is not as elegant and concise as

a word-grammar but it is expressive enough, and even more efficient when dealing with this kind of problems. Our mechanism supports the following two extra features:

- **bans** that can be stated together with a continuation class; they are used to express the set of continuation classes forbidden further along in the word-form (from the lexical entry defined with this restricted continuation class).

  ```
  bait       (PERTSONA - LA - N)
  ```

  This states that among the morphemes that follow the verb prefix *bait* in the word-form, those belonging to the continuation class *PERTSONA* are to be allowed but also that further on in the word no morphemes belonging to the continuation classes *LA* or *N* will be accepted. It always reduces the amount of continuation morphemes which can be linked after the morpheme.

- **continuation class-tree:** the lexicon builder has the possibility of changing the set of allowed continuation morphemes for a given one —so the amount of legal continuation morphemes can be reduced or incremented—, by means of making explicit these morphemes through different segments in the word-form; this explicitation is done by giving a parenthesized expression representing a tree. This mechanism improves the expressiveness of the formalism providing it with the additional power of specifying constraints to the set of morphemes allowed after the lexicon entry, stating in fact a continuation "path" —not restricted to the immediate morpheme— which makes explicit that set in a conditioned way.

  Long-distance dependency cases are found in the verb finite form instances above: the initial morpheme *na-* (absolutive, first person, present tense) allows dative morphemes corresponding to the third person after the morpheme *tzai* (root) but not those corresponding to the first person. Analogously the theoretically possible *hatzain** is not grammatical in Basque because it combines two second person morphemes in absolutive and dative cases. The continuation corresponding to *na* can be stated as follows:

  ```
  na    (KI (DAT23 (N_KE)), TZAI (DAT23 (LAT)))
  ```

  which specifies two alternative continuation "paths" allowed after this morpheme: the one including the morphemes in the continuation class *KI* and that which includes those in the continuation class *TZAI*. In both cases *DAT23* restricts the set of morphemes potentially permitted as continuation of those in *KI* or *TZAI*, allowing only the 2nd and 3rd person dative morphemes. Without this extension of the formalism, it would be possible to do it by storing repeatedly the morpheme *tzai* in two or more different lexicons, but this is not very useful when the distance between dependent morphemes is longer.

3.4 The Rules

Twenty four two-level rules have been defined to express the morphological, phonological and orthographic changes between the lexical and the surface levels that appear when the morphemes are combined. Given that suppletion cases are rare in Basque, phonemically unrelated allomorphs of the same morpheme are included in the lexicon system as separated entries. No rules deal with these phenomena. The rules are applied to express three types of realizations: adding or removing a character, or alternation of a character from the lexical to the surface level. These basic transformations can be combined. Although the whole set of rules is shown in the first appendix, three of the rules[2] will be explained here. Most of the rules are composite rules because the specified changes are normally obligatory.

voicing k

If place names ended in nasal consonant or any morphemes ended in surface *n* are combined with affixes beginning by epenthetical *e* and lexical *ko* this *k* is voiced to *g* when the epenthetical *e* is lost. This rule has interaction with another one which manages the mapping of the epenthetical e.

"description: voicing k"

```
k:g <=> [ Nasl %%: | :n ] MB (E:0) _ o ;
                   ! *usurbil%+Eko:*usurbilgo
                   ! *usurbil%+Eko:*usurbileko
                   ! egiN+ko:egingo
                   ! hemen+ko:hemengo
```

losing t

The lexical *t* is lost in these cases:

- at the end of a morpheme when the next one begins by unvoiced occlusive, nasal consonant or *h*.
- as part of an africated bigram at the end of a morpheme in some combinations.

"description: losing t"

```
t:0 <=> _ MB [ :ExpUnv | Nasl | h ] ;
          _ Silb %:0 MB E:0 ExpUnv ;
          _ Silb MB t ;
          n _ Silb MB k ;
                   ! bait+gara:baikara
                   ! *zarautz%+Eko:*zarauzko
```

---

[2] The syntax of the rules is taken from the compiler of Xerox (Karttunen & Beesley, 92).

```
                        ! utz+te:uzte
                        ! jantz+te:janzte
                        ! etxe+rantz+ko:etxeranzko
```

losing h

When a verb root beginning by *h* is linked to the prefix *beR* —equivalent to the prefix *re* in English— the *h* is lost.

"description: losing h"

```
  h:0 <=> R: MB _ ;
                        ! beR+hasi:berrasi
```

3.5 Results

Table 2 shows the output of the analysis of the sentence *"Eta gauza aundirik ekartzerik ez zuen izan" (And he/she could not bring anything important)*. *aundirik* is not the standard form (the correspondent standard one is *handirik*) and it is not analysed.

```
  ((form "*eta")
    ((anal 1)
     ((lemma "etA")((POS LINK))))
  )
  ((form "gauza")
    ((anal 1)
     ((lemma "gauza")((POS VERB))))
    ((anal 2)
     ((lemma "gauzA")((POS NOUN))))
    ((anal 3)
     ((lemma "gauzA")((POS NOUN)))
     ((morph "a")((POS DEC)(CAS ABS)(NUM S)(DET DEF))))
  )
  ((form "aundirik")
  )
  ((form "ekartzerik")
    ((anal 1)
     ((lemma "ekaR")((POS VERB)))
     ((morph "tzerik")((POS REL)(REL KONP))))
    ((anal 2)
     ((lemma "ekaR")((POS VERB)))
     ((lemma "tze")((POS ASP)(DERIV NOUN)))
     ((morph "Rik")((POS DEC)(CAS PAR)(DET UNDEF))))
  )
  ((form "ez")
    ((anal 1)
     ((lemma "ez")((POS ADV))))
    ((anal 2)
     ((lemma "ez")((POS NOUN))))
  )
  ((form "zuen")
    ((anal 1)
       ((lemma  "zuen")((POS  AUXV)(MD_TN  B1)(P_ABS  3)(P_ERG  3)(ROOT
*edun))))
    ((anal 2)
     ((lemma "zu")((POS PRON)))
     ((morph "eM")((POS DEC)(CAS GEN)(NUM P)(DET DEF))))
    ((anal 3)
```

```
       ((lemma  "zuen")((POS  AUXV)(MD_TN  B1)(P_ABS  3)(P_ERG  3)(ROOT
*edun)))
     ((morph "En")((POS REL)(REL RELAT))))
    ((anal 4)
       ((lemma  "zuen")((POS  AUXV)(MD_TN  B1)(P_ABS  3)(P_ERG  3)(ROOT
*edun)))
     ((morph "En")((POS REL)(REL IND_QUE))))
  )
  ((form "izan")
    ((anal 1)
     ((lemma "izaN")((POS VERB))))
    ((anal 2)
     ((lemma "izaN")((POS VERB)))
     ((morph "0")((POS ASP)(MOD PART))))
  )
```

Table 2 An example

In order to evaluate the coverage of the analyser we tested it with several corpora[3] and the results of two of the texts —*text1* a text of a magazine where many foreign names appear and *text2* a text about philosophy— are shown in Table 3 with two figures for each concept, one considering all the word tokens in the text (*corpus*) and other one considering only the different words (*list*).

Text	words	faults	hits(%)
1a.-Text1 (corpus)	4.864	379	**92,2**
1b.-Text1 (list)	2.607	307	**88,2**
2a.-Text2 (corpus)	2.343	95	**95,9**
2b.-Text2 (list)	1.429	85	**94,1**

Table 3 Texts for test

[3] The corpora are obtained from UZEI where have been stored in the project EEBS —Systematic compilation of the current Basque— (Urkia & Sagarna, 91).

Concept	1b	2b	total
Unknown words.	307 (%100)	85 (%100)	392 (%100)
A.-Non-standard use	101 (%32,9)	28 (%32,9)	129 (%32,9)
B1.-Loan-words	31 (%10,1)	2 (%2,4)	33 (%8,4)
B2.-Out of lexicon	68 (%22,1)	16 (%18,8)	84 (%21,4)
B3.-New derivatives	33 (%10,7)	13 (%15,3)	46 (%11,7)
B4.-Foreign words	39 (%12,7)	14 (%16,5)	53 (%13,5)
C.-Errors	30 (%9,8)	10 (%11,8)	40 (%10,2)
D.-Others	5 (%1,6)	2 (%2,4)	7 (%1,8)

Table 4 Causes of the faults

As the coverage was not satisfactory enough, we tried to find the causes of the faults sorting them into different sets (see Table 4):

A) non-standard uses or linguistic variants: due to the recent standardisation and the widespread dialectal use of Basque the use of non-standard forms is quite wide. These non-standard forms were not recognized by the system because we wanted the morphological processor to generate only standard forms and, as we will explain below, we wanted the spelling checker to be be able to detect them.

B) entries not in the lexicon because of other reasons: foreign words, loan-words, new unpredictable derivatives, and others.

C) errors in the texts and other problems.

Keeping in mind these figures, it was necessary to manage non-standard uses and forms whose lemmas are not in the lexicon if we wanted to develop a comprehensive analyser. This management is explained in next section.

As to the issue of speed of our processor can analyse two or three words per second, amounts similar to others using PC-Kimmo (Antworth, 90) but not enough for some on-line processes. As we show below, the use of lexical transducers is a good alternative to improve the speed.

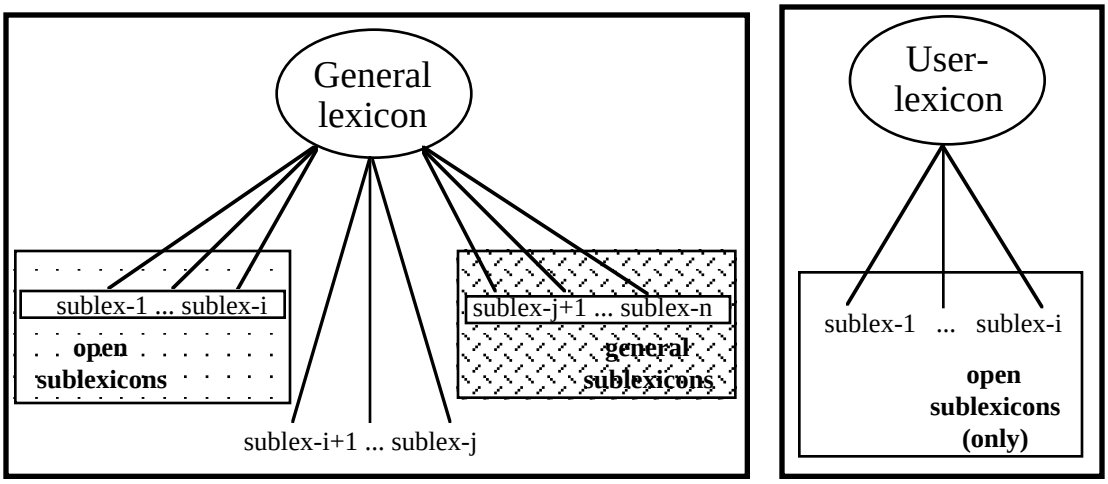
4. Increasing the Coverage and the Robustness

In this section we explain three extra-features of our analyser that have been introduced in order to improve the coverage: the management of a user-lexicon, the treatment of linguistic variants and the analysis of unknown words.

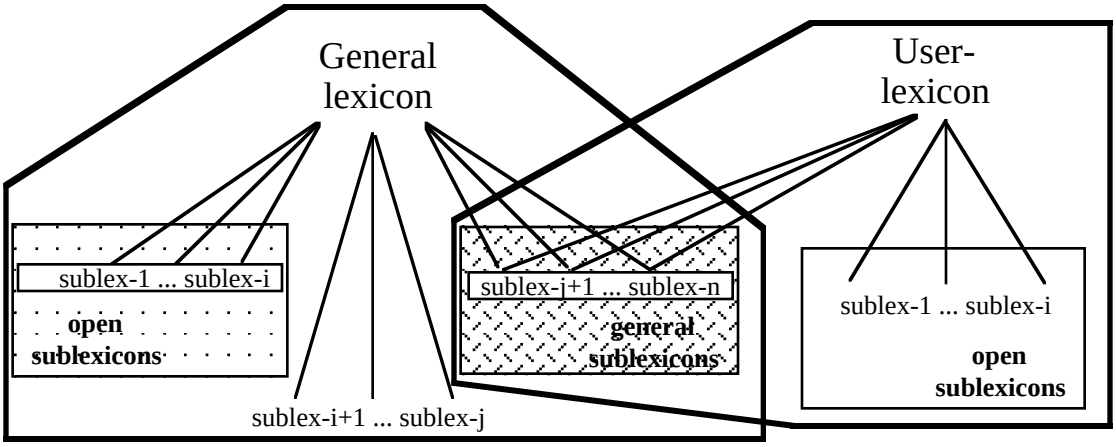
4.1 Using User-Lexicon

Analysing forms whose roots are not in the general lexicon is possible if the user can update the lexicon, —the general lexicon or a personal or user-lexicon. The second option is more flexible and it is this which we use. User-lexicons can be interactively enriched by means of a specially designed human-machine dialogue which allows the system to acquire the internal features of each new entry (sublexicon, continuation class, and selection marks). It is very important to notice the necessity of a suitable interface for lexical knowledge acquisition when it comes to managing with precision the inclusion of new lemmas in the user's own dictionary. Without this interface morphological and morphotactical information essential to the checker would be left unknown hence, no inflected forms could be accepted. At present, the system acquires information from the user about part of speech, subcategorization for nouns —person or place names, mainly— and some morphonological features like final hard-or-soft r distinction. So, the user, giving to the system several answers makes the correct assignment of continuation class and selection marks to the new lemma possible. In this way, open class entries may be accepted and adequately treated. Entries belonging to other classes may also be entered but no flexion of them will be recognized. This ability to deal correctly with new lemmas requires, in turn, certain grammatical knowledge from the user.

The management of the user-lexicon is done defining some sublexicons as open and multiplexing the use of these open sublexicons between general and user-lexicons during the different steps of the analysis (see Fig. 5). Six sublexicons have been defined as open and can be updated in the user-lexicon by means of the interface.



(A) General lexicon and user-lexicon (in external files)



(B) Managing the user-lexicon

Fig. 5 Structure of the user-lexicons

4.2 The Analysis of Linguistic Variants

Because of the recent standardisation and the widespread dialectal use of Basque, the standard morphology is not enough to offer good results when analysing corpora.

Three types of linguistic variants are distinguished: morpheme variants —i.e. *haundi* is used instead of standard *handi* (big)—, morphotactical variants —i.e. the standard declension of *batzu* (someone) is plural but it is often declined as indeterminate— and morphonological variants or regular non-standard changes —i.e. the use of the *h* was controversial and it is not yet well known.

The treatment of these variants has been carried out by means of an additional two-level subsystem (Aduriz et al., 93), thus increasing the coverage of the morphological processor.

This subsystem is also used in the spelling corrector to manage competence errors and has two main components:

1) New morphemes linked to the corresponding correct ones. They are added to the lexical system and they describe particular variations, mainly dialectal forms. More than 1000 non-standard morphemes —mainly dictionary entries— have been included in this subsystem.

2) New two-level rules describing the most likely regular morphonological changes that are produced in the variations. These rules have the same structure and management than the original ones. Eighteen new rules have been defined to cover the most common competence errors.

The non-standard analyses are rejected if there are standard ones. When different non-standard analyses are obtained there is a disambiguation process that prefers concrete analysis (morpheme or morphotactical variants) to general ones (morphonological variants) and, among these analyses, those with less non-standard morphonological rules are applied.

Results

Using the list of unknown words referenced at Table 3 and Table 4 we tested our non-standard subsystem and we have concluded that it is possible to analyse correctly about the 80% of the linguistic variants (see Table 5).

Concept	**1b**	**2b**	**total**
Unknown words.	307	85	392
Non-standard words	101 %100	28 %100	129 %100
Analysed	85 **%84,2**	22 **%78,6**	107 **%83**

Table 5 Evaluation of the analysis of linguistic variants

4.3 The Analysis of Unknown Words

Based on an idea used in speech synthesis (Black et al., 91), a two-level mechanism for analysis without lexicon was added to increase the robustness of the analyser.

This mechanism of treatment of unknown words has two main components: 1) generic lemmas represented by "*??*" —one for each possible open category or subcategory— which are stored along with the general affixes in a small two-level lexicon, and 2) two additional rules that express the relationship between the generic lemmas at the lexical level and any acceptable lemma of Basque, which are combined with the standard rules.

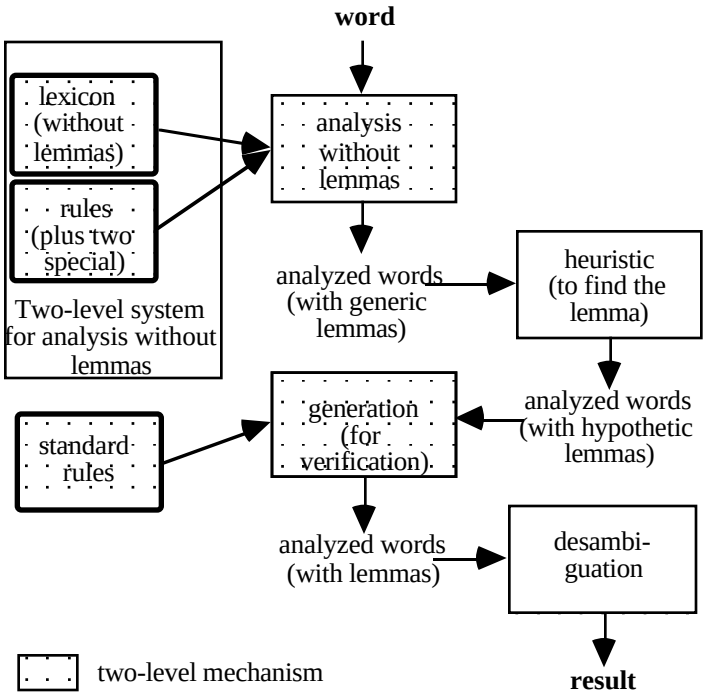


Fig. 6 Analysis of unknown words

Some standard rules have to be modified because surface and lexical level are specified and in this kind of analysis the lexical level of the lemmas is made out the generic lemmas. The same two-level mechanism is used to analyse the unknown forms and the obtention of at least one analysis is guaranteed.

As a result of the analysis generic lemmas and concrete affixes are obtained. A heuristic is responsible for finding concrete possible lemmas instead of the generic ones, and standard generation helps in this process because it is always possible to verify that the combination of hypothetical lemmas and affixes is right. In order to eliminate the great number of ambiguities in the analysis, a local disambiguation process —a function of the length and the last characters of the hypothetical lemmas— is carried out (Fig. 6).

4.4 Results

Figures about the precision of the analyser are given in Table 6 for the same texts mentioned above. We can conclude that it is a high-coverage and robust analyser. This analyser is a basic tool for current and future work on automatic processing of Basque and it is intensively used in the process of spelling correction that we describe below.

Concept	Text 1	Text 2	Total
Different words (list)	2.607	1.429	4.036
Unknown words in standard analysis	307 %12	85 %6	392 %10
Linguistic variants Recognised variants	101 85 (%84)	28 22 (%79)	129 107 (%83)
Errors after all analyses **Precision**	21 **%99,2**	4 **%99,7**	25 **%99,4**

Table 6 Precision of the analyzer

5 Improving Morphological Analysis Using Lexical Transducers

A lexical transducer (Karttunen et al., 92) (Karttunen, 94) is a finite-state automaton that maps inflected surface forms onto lexical forms, and can be seen as an evolution of two-level morphology where:

- Morphological categories are represented as part of the lexical form. Thus, it is possible to avoid the use of diacritics.
- Inflected forms of the same word are mapped into the same canonical dictionary form. This increases the distance between the lexical and surface forms. For instance *better* is expressed through its canonical form *good* (*good+COMP:better*).
- Intersection and composition of transducers is possible (see Kaplan and Kay, 94). In this way the integration of the lexicon —the lexicon will be another transducer— in the automaton can be solved and the changes between lexical and surface level can be expressed as a cascade of the two-level rule systems (Fig. 7).

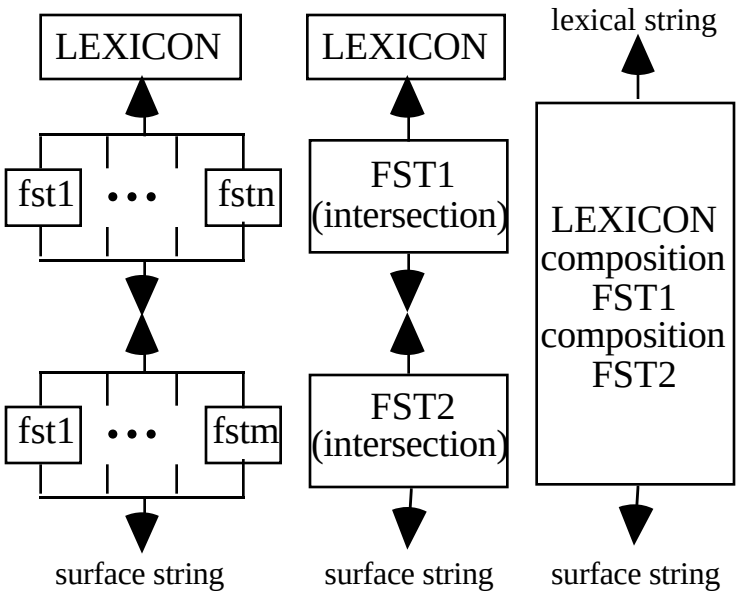


Fig. 7 Lexical transducers (from Karttunen et al., 92)

In addition, the morphological process using lexical transducers is very fast —thousands of words per second— and the transducer for a whole morphological description can be compacted in less than 1Mbyte. We are using the tools developed in Xerox to build lexical transducers (Karttunen & Beesley, 92) (Karttunen, 93). Uses of lexical transducers are documented by Chanod (Chanod, 94) and Kwon (Kwon & Karttunen, 94). We have used lexical transducers in order to improve both the linguistic description and the speed of the morphological analysis (Alegria *et al*., 95).

The conversion of our description into a lexical transducer was done through the following steps:

a) Canonical forms and morphological categories were integrated into the lexicon from the lexical database.
b) Due to long distance dependencies among morphemes, which could not be resolved in the lexicon, two additional rules were written to ban some combinations of morphemes. These rules can be put in a different rule system near to the lexicon without mixing morphotactics and morphonology.
c) The standard rules could be left unchanged but were changed in order to replace diacritics with morphological features, so doing a better description of the Basque morphology.

The resultant lexical transducer is 500 times faster than the original system.

Conclusions

A two-level formalism based morphological processor for Basque has been designed in an incremental way. It has three main modules: the standard analyser, the analyser of linguistic variant, and the analyser without lexicon. User-lexicons can be interactively enriched with new entries enabling the analyser to recognize all the possible flexions derived from them. The analyser is very flexible, has a wide coverage, and is a basic tool for current and future work on automatic processing of Basque. Using lexical transducers for our analyser we have improved both the speed and the description of the different components of the tool. The results have been described in detail to explain the quality, scale and precision.

Acknowledgements

This work has had partial support from the Economy Department of the Local Government of Gipuzkoa and from the Education Department of the Government of the Basque Country. We would also like thank to Xerox for letting us use their tools, and to Ken Beesley and Lauri Karttunen for their help using tools and designing lexical transducers. Thanks to Eneko Agirre for help in with the implementation of extended continuation classes and to all the members of our resarch team. We are in debt with Hitesh Patel for his help in the English version of this paper.

References

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. (1995). Different issues in the design of a lemmatizer/tagger for Basque. *"From text to tag" Workshop*, SIGDAT, EACL.

Agirre E., Alegria I., Arregi X., Artola X., Diaz de Illarraza A,. Maritxalar M., Sarasola K., Urkia M. (1992). XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology, *Proc.of the Third ANLP*, 119-125.

Agirre E., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Insausti J.M., Sarasola K. (1995). Different issues in the design of a general-purpose Lexical Database for Basque. *NLDB'95 Workshop*.

Alegria I. (1995). *Euskal morfologiaren tratamendu automatikorako tresnak*. Ph.D. Thesis. In Basque.

Alegria I., Artola X., Sarasola K. (1995). *Improving a robust morphological analyser using lexical transducers*. Recent advances in Natural Language Processing. Bulgary.

Antworth E.L. (1990). *PC-KIMMO: A two-level processor for morphological analysis*. Occasional Publications in Academic Computing, No. 16, Dallas, Texas.

Barton G.E. (1986). Computational Complexity in two-level Morphology, *ACL Proceedings, 24th* Annual Meeting.

Bear J. (1986). A morphological recognizer with syntactic and phonological rules. *Proc. of COLING´86*, 272-276.

Black A., van de Plassche J., Williams B. (1991). Analysis of Unkown Words through Morphological Descomposition. *Proc. of 5th Conference of the EACL*, vol. 1, 101-106.

Carter D. (1995). Rapid development of morphological descriptions for full language processing system. *Proc. of EACL´95*.

Chanod J.P. (1994). *Finite-state composition of french verb morphology*. Xerox MLTT-005.

Euskaltzaindia (1985). *Euskal Gramatika: Lehen urratsak (I, II, III eta IV)*. Euskaltzaindia, Bilbo.

Kaplan R. M. and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguistics*, vol.20(3), 331-380.

Karttunen L., Kaplan R.M., Zaenen A. (1992). Two-level morphology with composition. *Proc. of COLING´92*.

Karttunen L. and Beesley K.R. (1992). *Two-Level Rule Compiler*. Xerox ISTL-NLTT-1992-2.

Karttunen L. (1993). *Finite-State Lexicon Compiler*. Xerox ISTL-NLTT-1993-04-02.

Karttunen L. (1994). Constructing Lexical Transducers, *Proc. of COLING´94*, 406-411.

Koskenniemi, K. (1983). *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics. Publications nº 11.

Oflazer K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, vol.9, No. 2, 137-148.

Ritchie G., S.G. Pulman, A.W. Black and G.J. Russell (1987). A Computational Framework for Lexical Description, *Computational Linguistics*, vol. 13, ns 3-4.

Ritchie G., A.W. Black, G.J. Russell and S.G. Pulman (1992). *Computational Morphology*.The MIT Press.

Solack A, Oflazer K. (1993). Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing*, vol.8, No. 3, 113-130.

Sproat R. (1992). *Morphology and Computation*. The MIT Press.

Trost H. (1990). The application of two-level morphology to non-concatenative German morphology, *Proc. of COLING-90*, Helsinki, vol.2 371-376.

Urkia M, Sagarna A. (1991). Terminología y Lexicografía asistida por ordenador. La experiencia de UZEI, *SEPLN*, vol 8.